Revolutionizing the NBA Scouting Industry

BAN A - Team 3

Vincent, Candela, Esteban, Mehul

SCOUTING COMPARISON

VS

Stephen Curry

Michael Beasley

CURRY'S HIGH SCHOOL RANKINGS

Rivals Rating ★★★☆☆

RANKINGS POSITION	NATIONAL	STATE
NO RATING	NO RATING	NO RATING
Rivals Rating	About	4.9 — 6.1

No Division 1 College Offers

2009 NBA Scouting Report:

- Stephs' explosiveness and athleticism are below standard.
- He's not a great finisher around the basket.
- He needs to considerably improve as a ball handler
- Stephen must develop as a point guard to make it in the league

Scouting Report
Stephen Curry

Scouting Report
Michael Beasley

BEASLEY'S HIGH SCHOOL RANKINGS

Rivals Ranking ★★★★★

RANKINGS

POSITION	NATIONAL	STATE
NO RATING	NO RATING	NO RATING

Rivals Rating — About 4.9 – 6.1

3 Division 1 College Offers

2008 NBA Scouting Report

- He averaged 20 points per game at Oak Hill Academy.
- Scored 64 points in a single game at Notre Dame School.
- "Most promising MVP player pick for 2008" by Miami Heat.

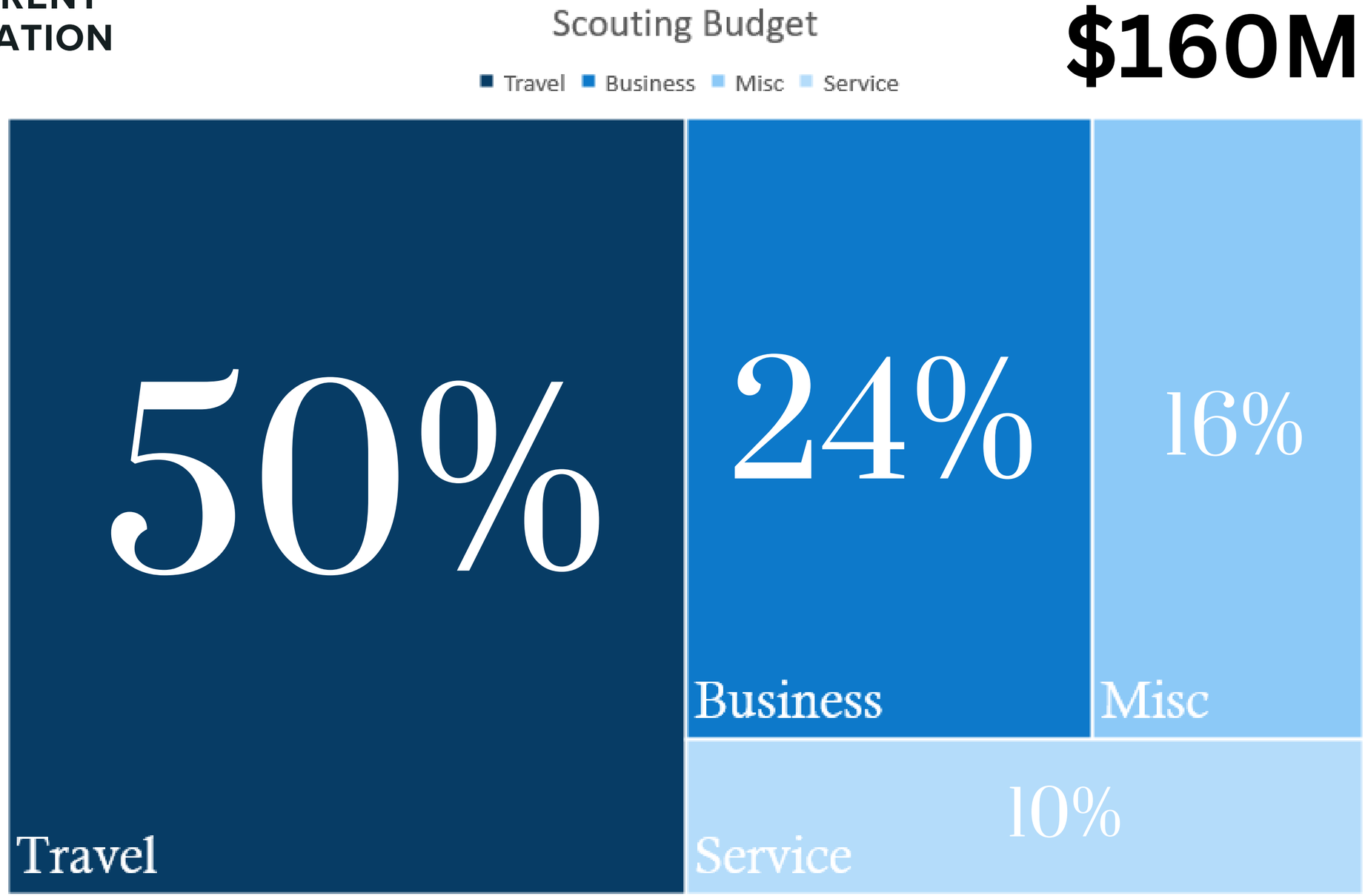
In Reality.

Championships & MVPs: 2015, 2017, 2018, 2022

In Reality.

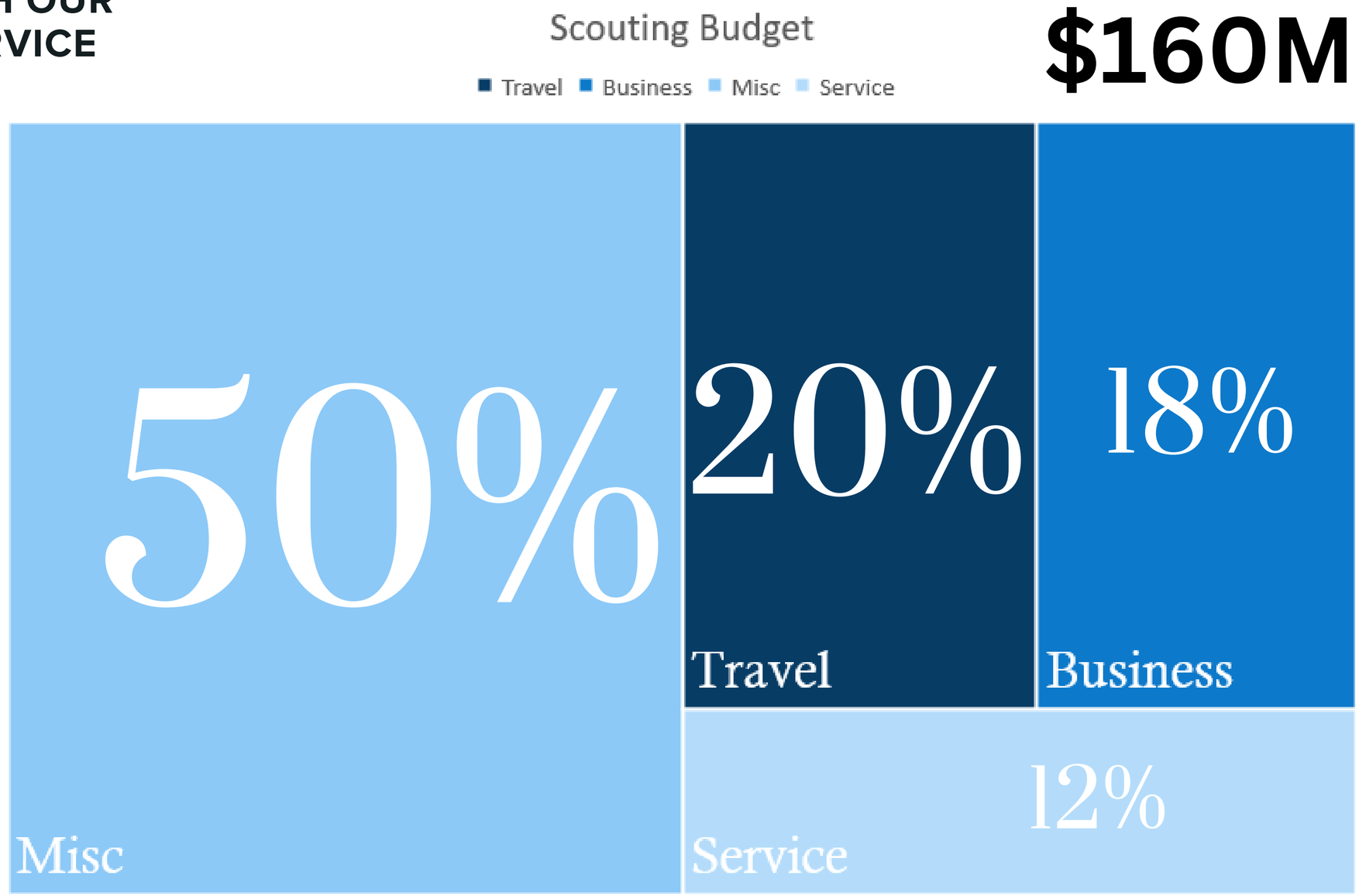
0 Championships & MVPs

CURRENT SITUATION



$160M

WITH OUR SERVICE



$160M

streamlining international travel and using our algorithm as a tool.

THE PROBLEM AT HAND

- Budgets are strained which **hinders optimal resource allocation** and financial flexibility.
- The existing accuracy level of MVP-caliber player predictions has higher false negatives, leading to **suboptimal decision-making** processes and compromising team performance.
- Difficulty in differentiating the team's brand and attracting top talent due to a lack of innovative marketing strategies

OUR PROPOSED SOLUTION

- Establish competitive advantage through enhanced player prediction accuracy, bolstering market positioning.
- Deliver budgetary freedom by minimizing false negatives in MVP-caliber player predictions, and **optimizing resource allocation.**
- Elevate MVP-caliber player predictions to ensure a superior edge in decision-making and team performance.

MACHINE LEARNING EXPLORATION

IMPORTING THE DATA

Before data cleaning

Df.shape

(14573, 32)

After data cleaning

Df.shape

(14573, 27)

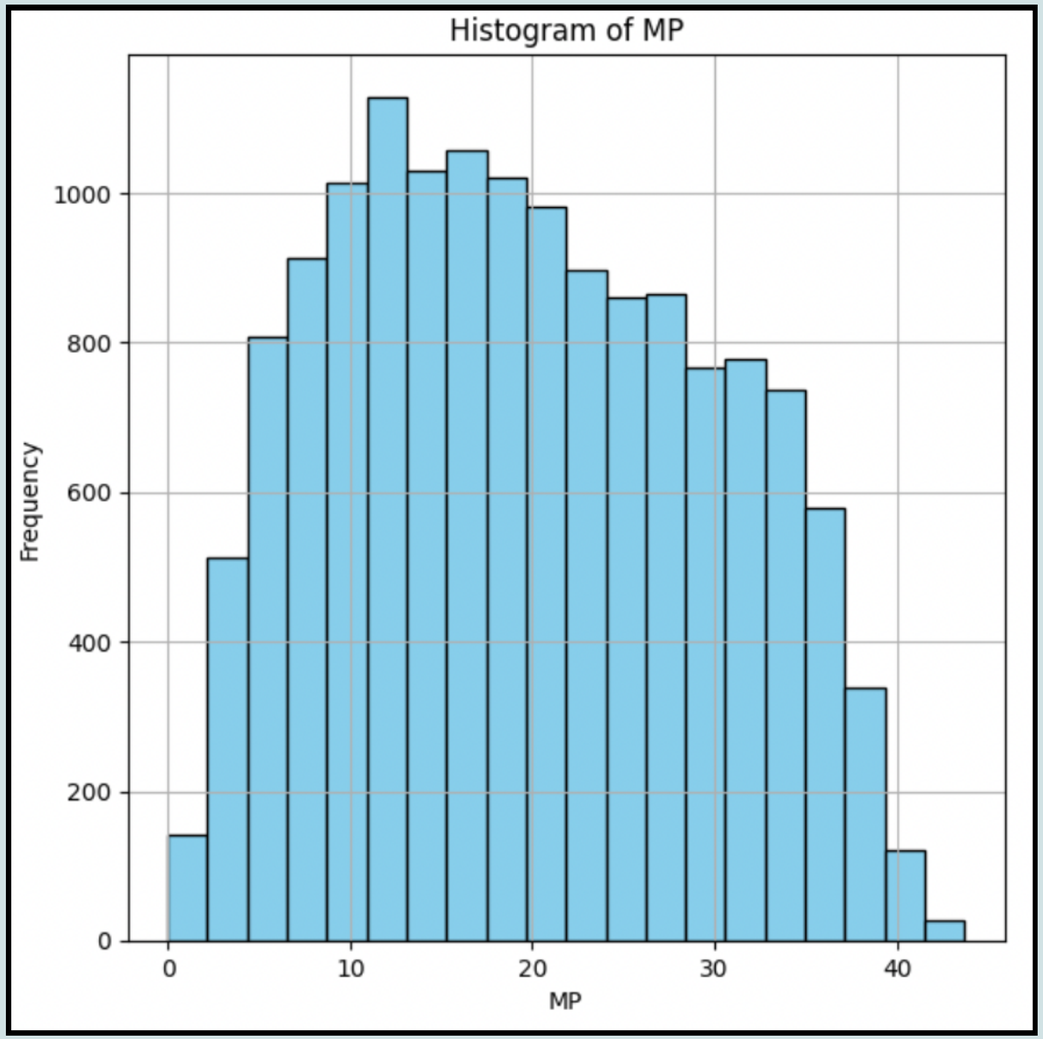
```
# Display the first few rows of the DataFrame to verify the data
print(df.head(5))
```

```
   Rk                Player Pos  Age   Tm   G  GS    MP   FG   FGA  ...  DRB  \
0   1   Mahmoud Abdul-Rauf  PG   28  SAC  31   0  17.1  3.3   8.8  ...  1.0
1   2    Tariq Abdul-Wahad  SG   23  SAC  59  16  16.3  2.4   6.1  ...  1.2
2   3  Shareef Abdur-Rahim  SF   21  VAN  82  82  36.0  8.0  16.4  ...  4.3
3   4       Cory Alexander  PG   24  TOT  60  22  21.6  2.9   6.7  ...  2.2
4   4       Cory Alexander  PG   24  SAS  37   3  13.5  1.6   3.9  ...  1.1

   TRB  AST  STL  BLK  TOV   PF   PTS    Season    MVP
0  1.2  1.9  0.5  0.0  0.6  1.0   7.3  1997-98  False
1  2.0  0.9  0.6  0.2  1.1  1.4   6.4  1997-98  False
2  7.1  2.6  1.1  0.9  3.1  2.5  22.3  1997-98  False
3  2.4  3.5  1.2  0.2  1.9  1.6   8.1  1997-98  False
4  1.3  1.9  0.7  0.1  1.3  1.4   4.5  1997-98  False

[5 rows x 32 columns]
```

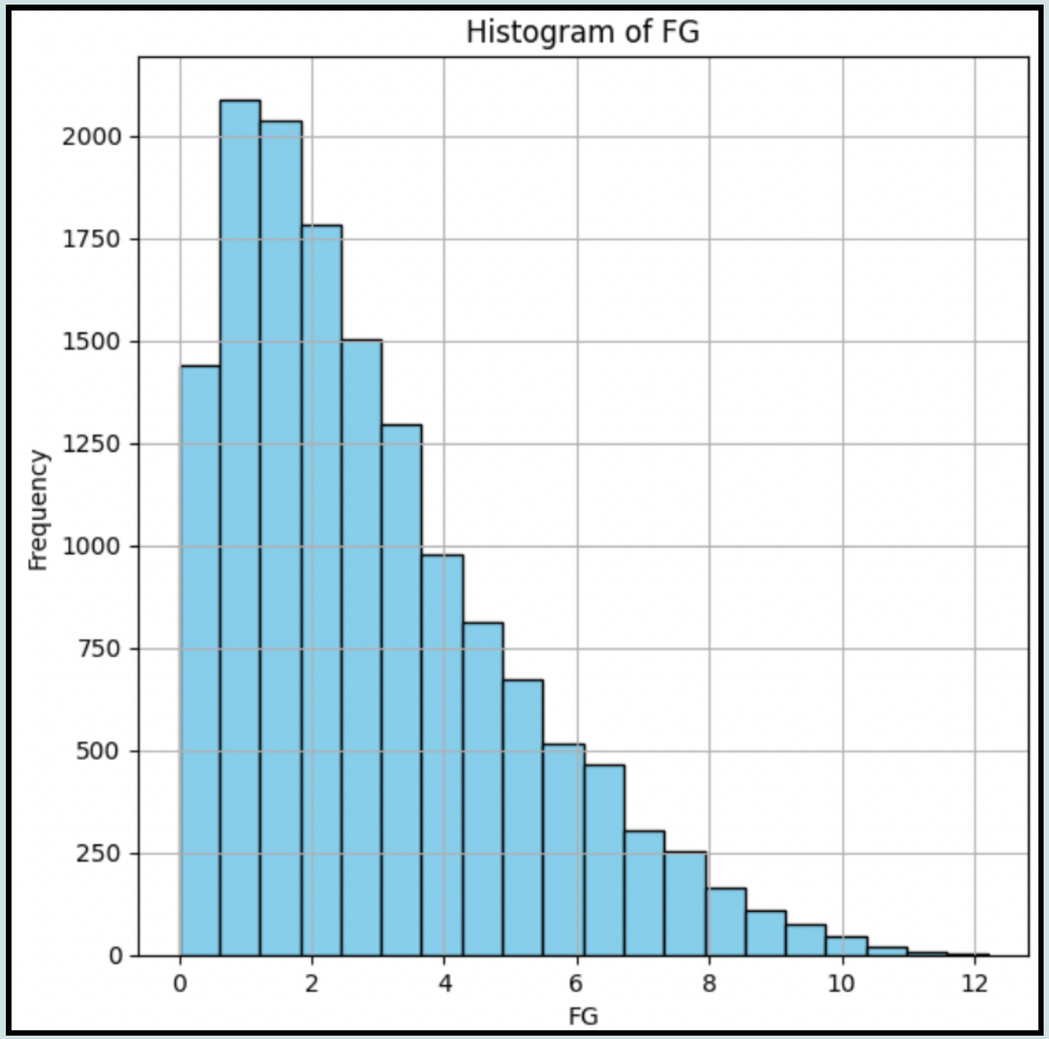
Dataset:
https://data.world/etocco/nba-player-stats

Exploratory Data Analysis

Matches Played

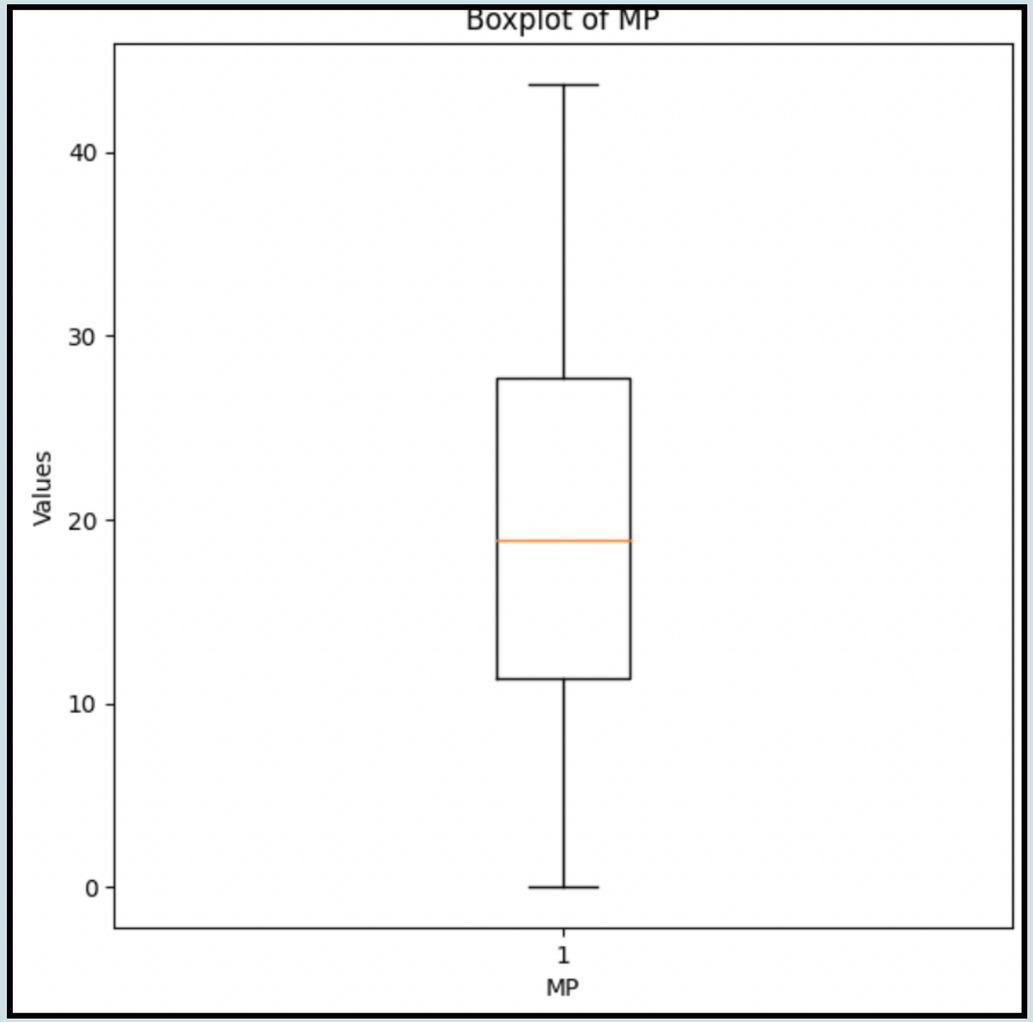


Field Goals

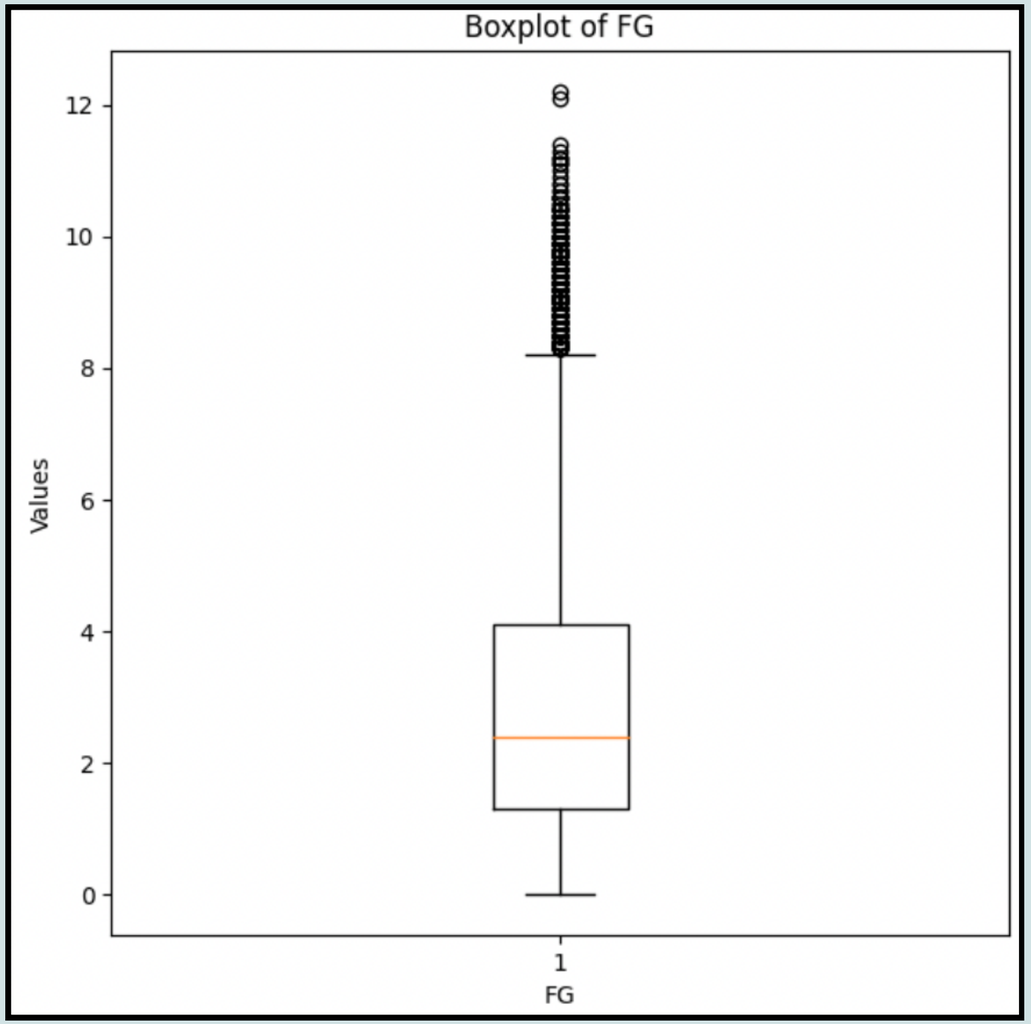


Exploratory Data Analysis

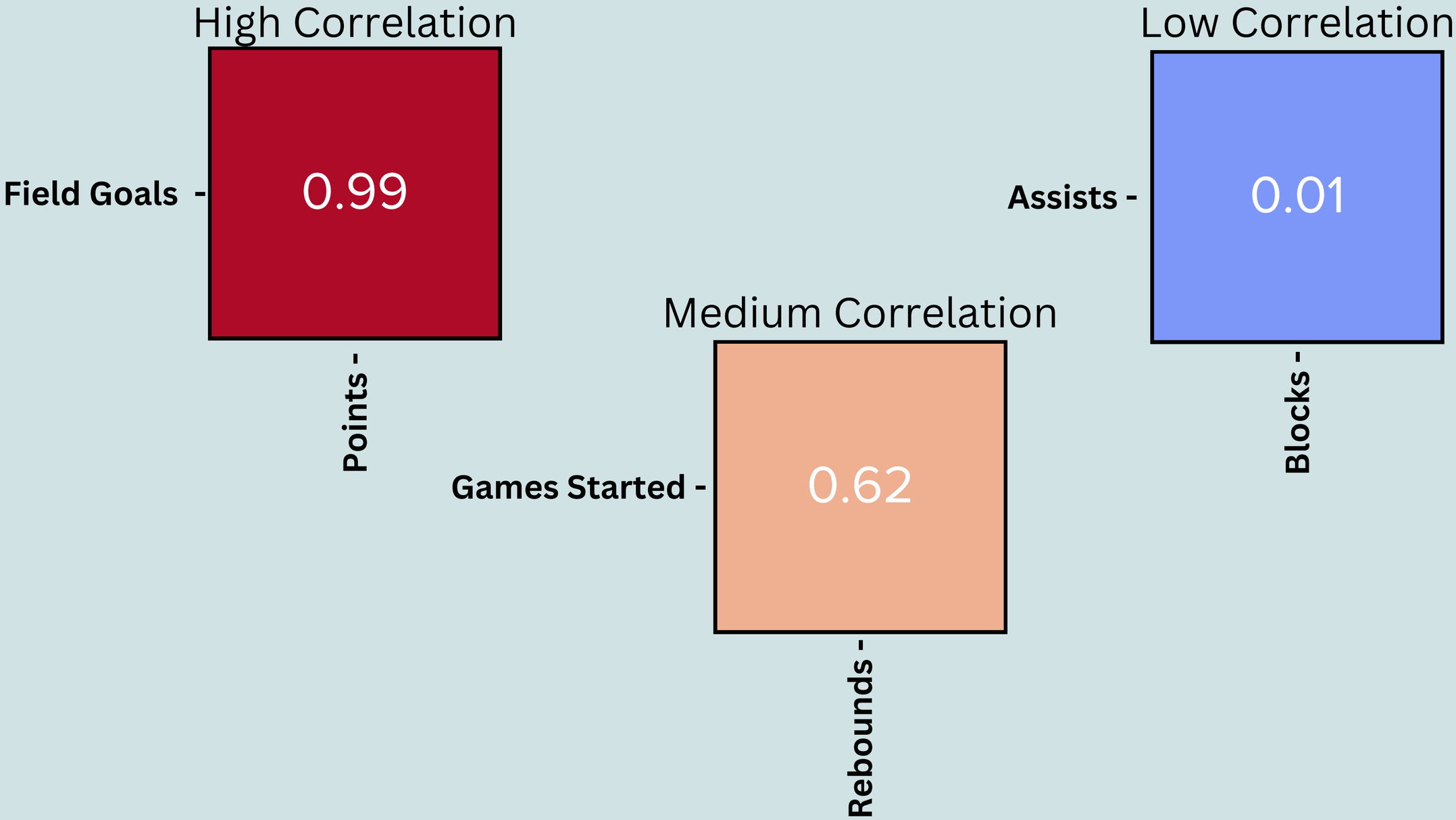
Matches Played



Field Goals

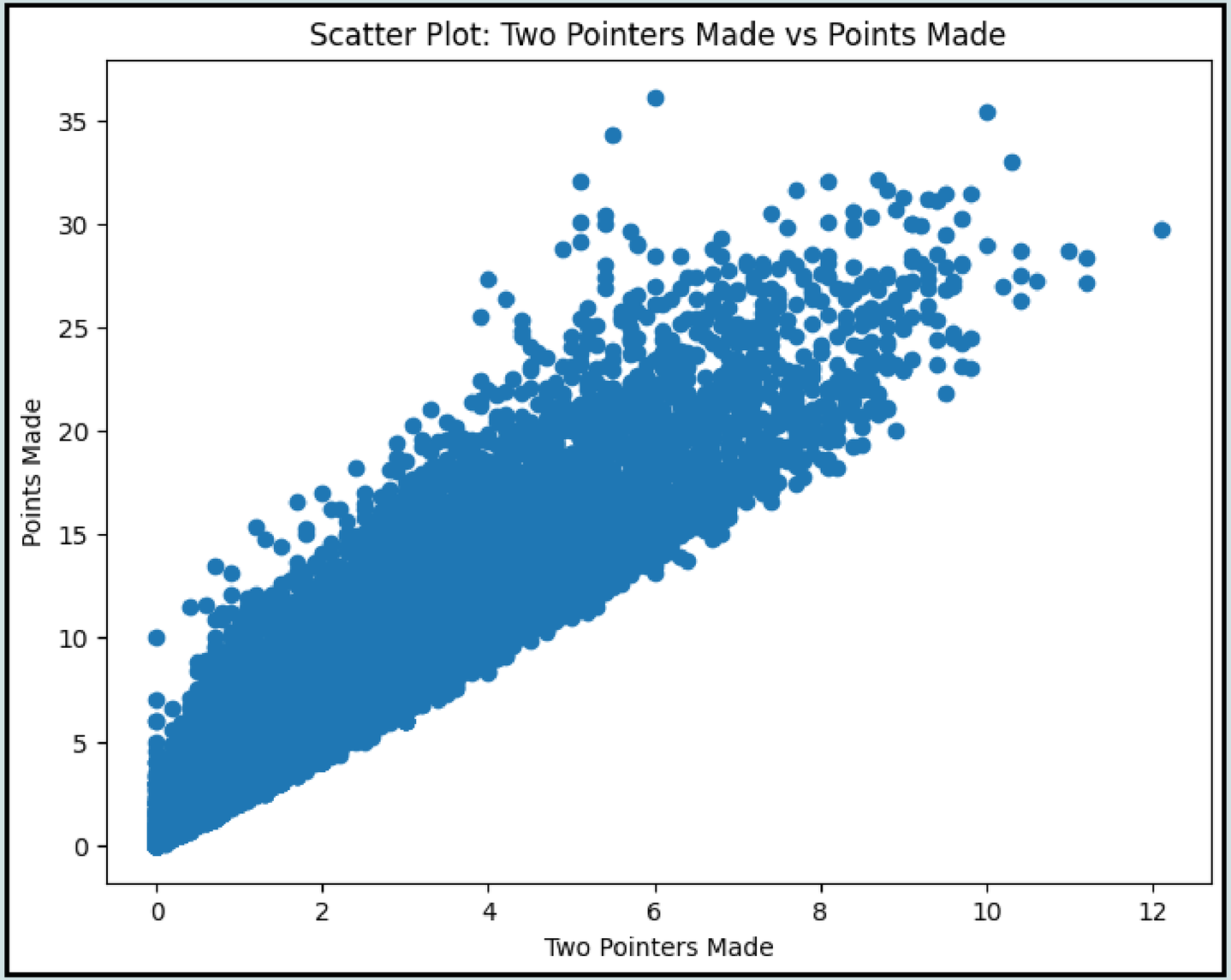


Correlation Matrix



Bivariate Data Analysis

Two pointers vs. Points made

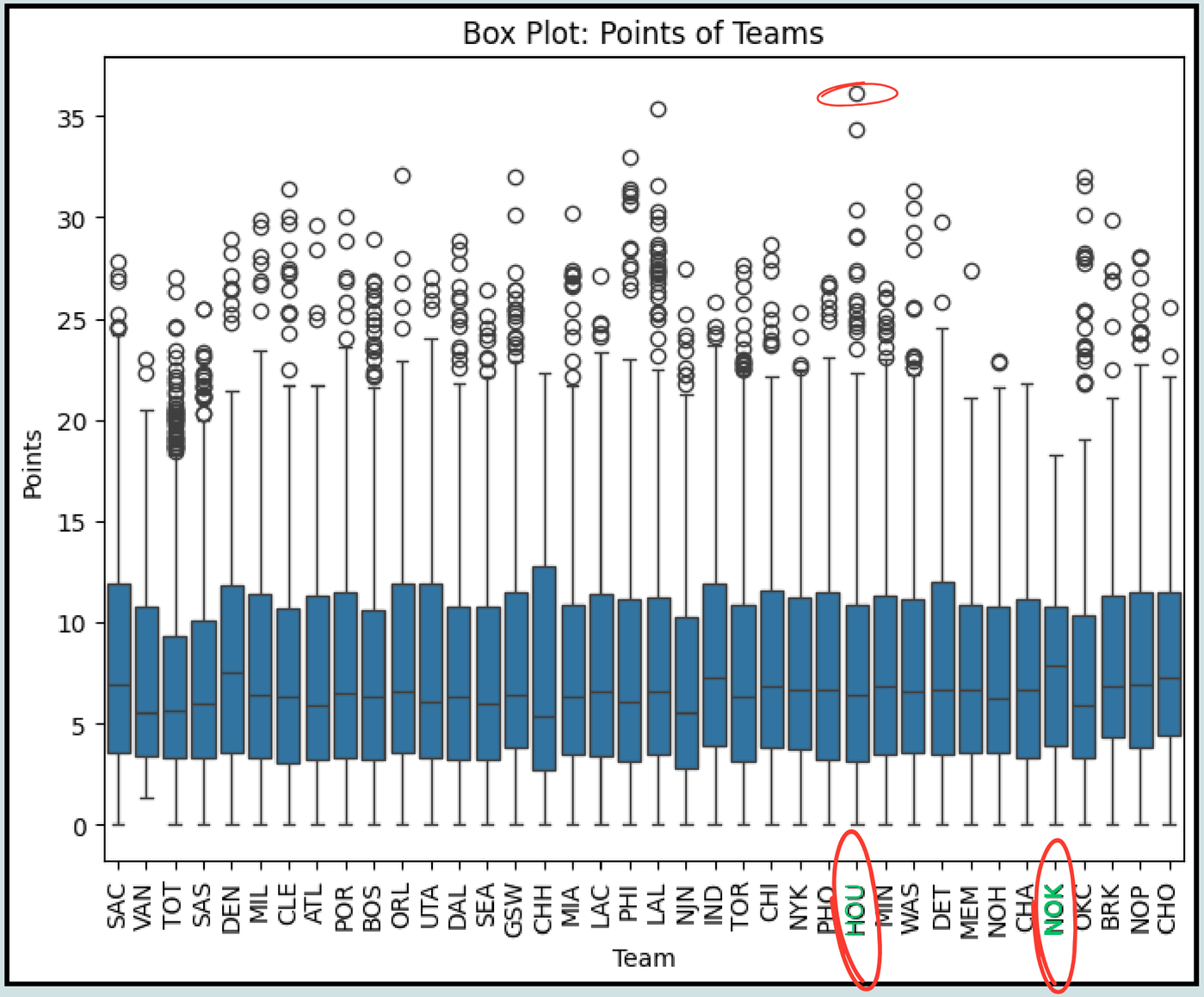


Conclusion:

- Two pointer make up large amount of the total points

Bivariate Data Analysis

Points distribution by player of a team

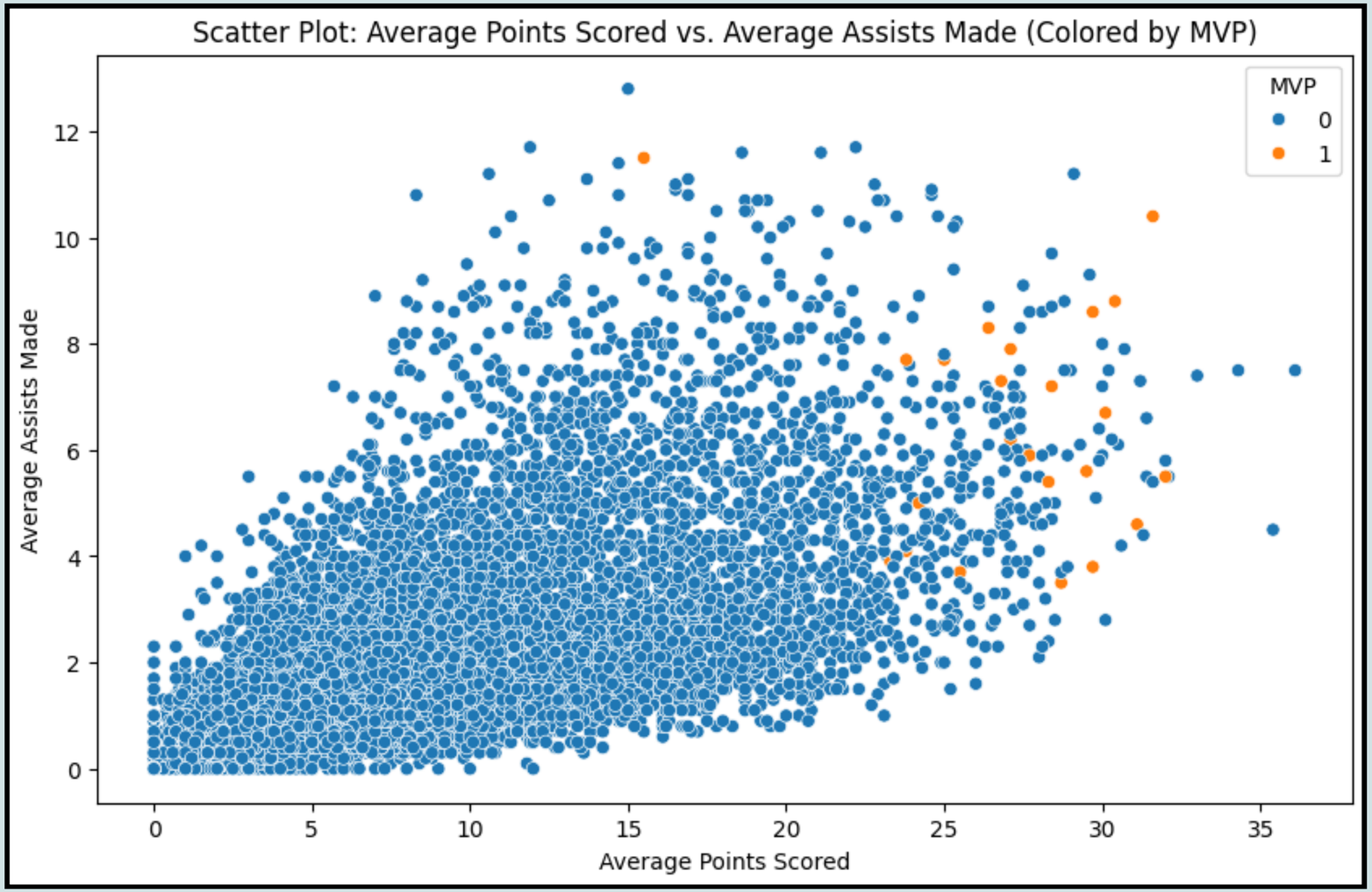


<u>Conclusion:</u>

- **Houston Rockets** has the <u>player</u> with the <u>most average points</u>
- **New Orleans Pelicans** has no player that stands out (lowest budget).

Multivariate Data Analysis

Average Points Scored vs Assists Made colored by MVP

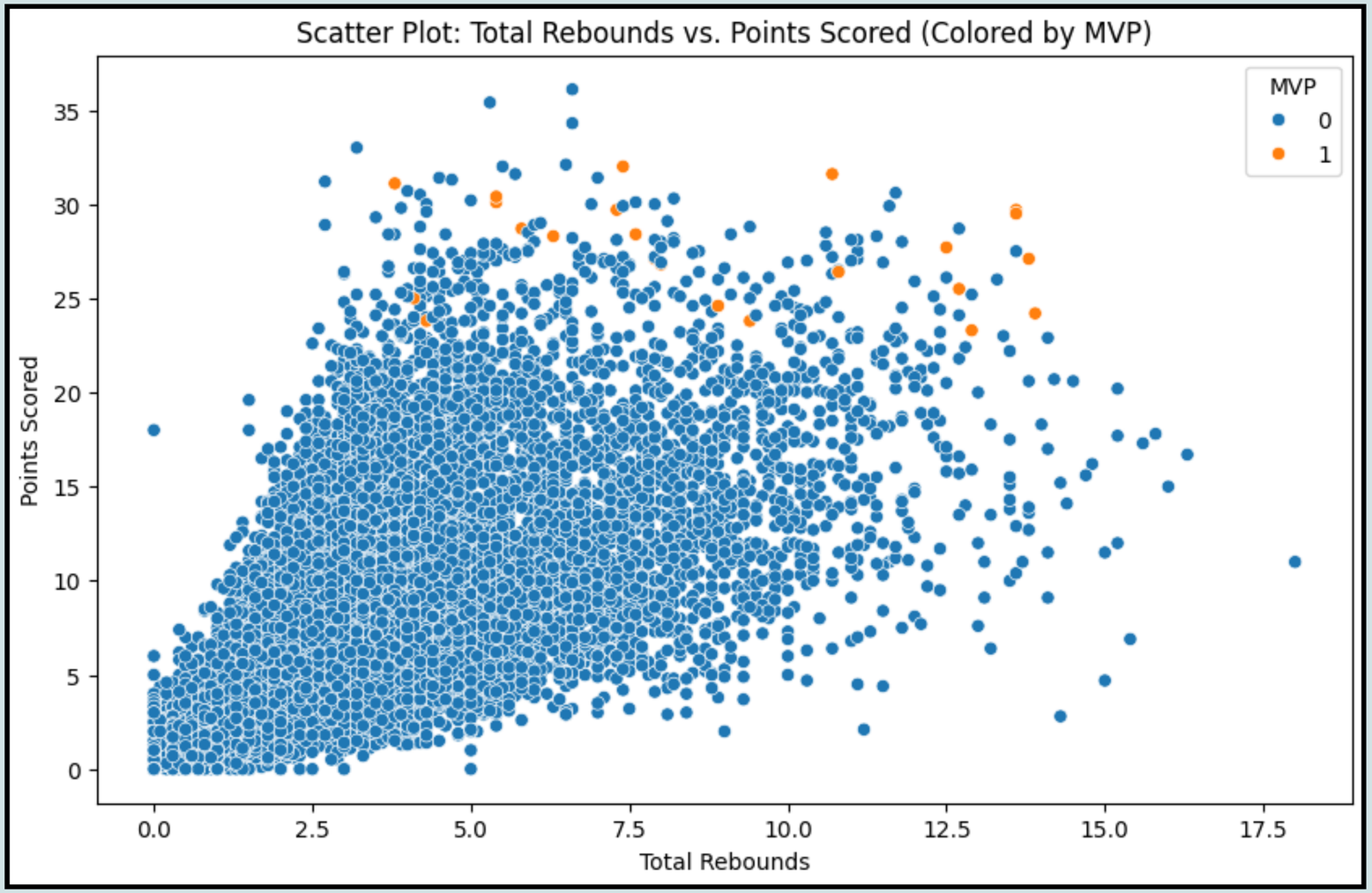


<u>Conclusion:</u>

- **<u>MVP's</u>** are outstanding at making **assists** and scoring **points.**

Multivariate Data Analysis

Total Rebounds vs Points Scored, colored by MVP

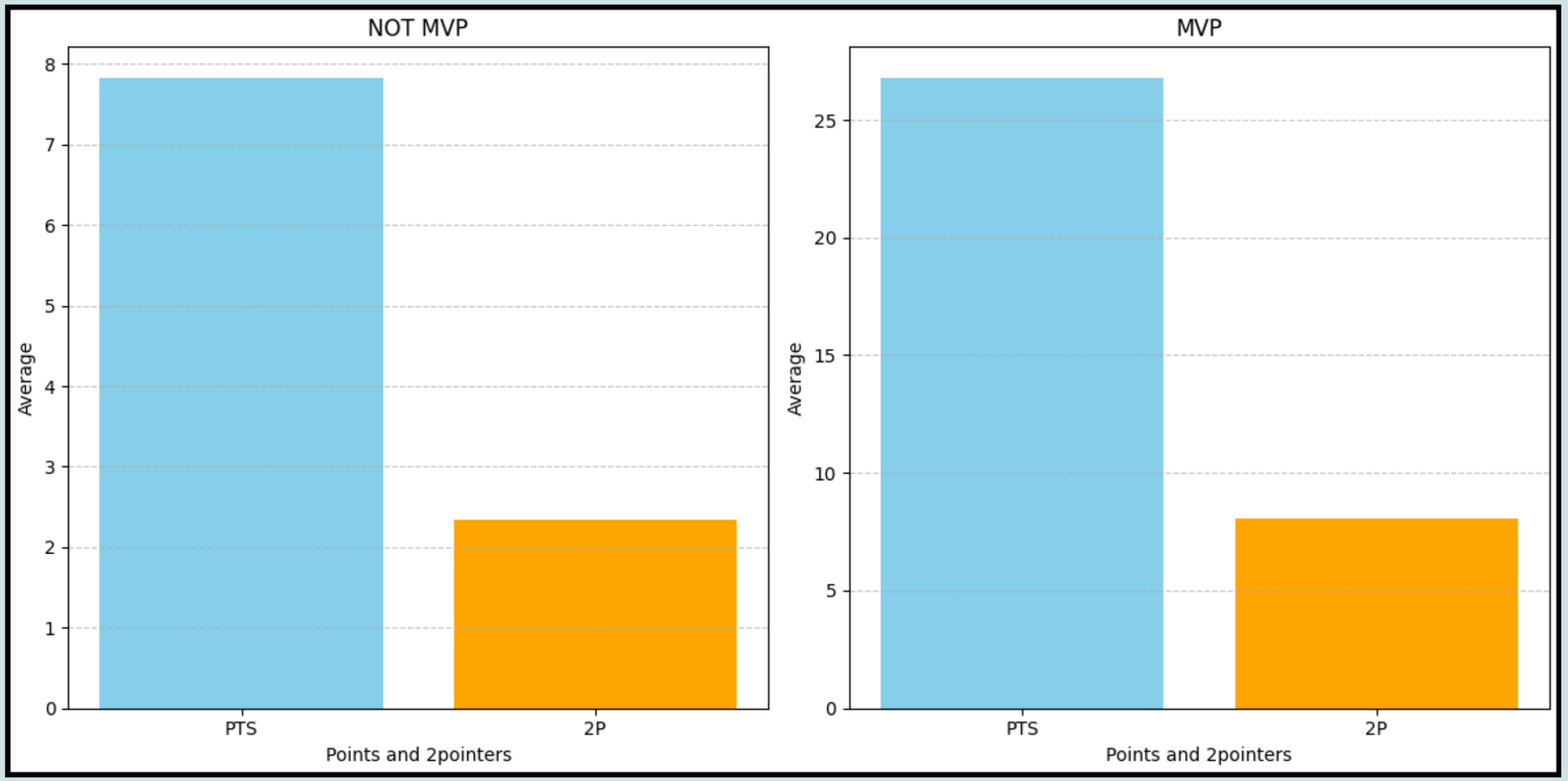


Conclusion:

- **MVP's** are outstanding on making **points.**
- **Rebounds** have **less affect** on the MVP distribution.

Hypothesis testing for MVP

MVP vs not MVP points and 2 pointers



Conclusion:

- There is a significant difference in the scale between **MVP's** and not **MVP's**

Data Cleaning and Scaling

Removing outliers and standardizing the data

```
# Select only the numerical features
numerical_features = df.select_dtypes(include=['float64', 'int64'])

# Initialize the scaler
scaler = StandardScaler()
```

We standardized the data instead of normalizing it due to the difference in scale.

We **didn't remove outliers** from this data set because the outliers in this case are the **MVP's.**

Feature Selection

Techniques for Selecting Features to Reduce Dimensionality

KBest

K = 5

Features:

- Field Goals
- 2 Pointers
- Free Throws
- Free Throw Attempts
- Points

KBest focuses on shooting performance

LASSO

Alpha = 0.6

Features:

- Points
- Personal Fauls
- Age
- Games Played
- Games Started

LASSO focuses on individual statistics

Decision Tree

Features:

- Matches Played
- Field Goals
- Assists
- Personal Fauls
- Total Rebounds

Decision Tree focuses on the 3 key metrics in basketball: Points, Assists, Rebounds

Interaction Terms

Interaction Terms to Add Relevant Features

Field Goal Efficiency

Field Goals made / Field Goals Attempted

A higher value of this interaction term indicates better shooting efficiency, reflecting a player's ability to convert field goal attempts into points effectively.

Rebounding Impact

Total Rebounds * (Offensive Rebounds + Defensive Rebounds)

It reflects the player's ability to secure possessions for their team and contribute to both offensive and defensive play.

Scoring Versatility

Total Points / (Field Goals + Three Pointers + Free Throws)

A higher value of this interaction term indicates a player's ability to score points efficiently from various scoring methods, showcasing their versatility and scoring prowess.

Feature Selection with Interaction Terms

KBest

K = 5

Features:

- Field Goals
- 2 Pointers
- Free Throws
- Free Throw Attempts
- Points

No Changes

LASSO

Alpha = 0.6

Features:

- Scoring Efficiency
- Free Throws
- Age
- Games Played
- Games Started

Scoring Efficiency Replaces Points Scored.

Decision Tree

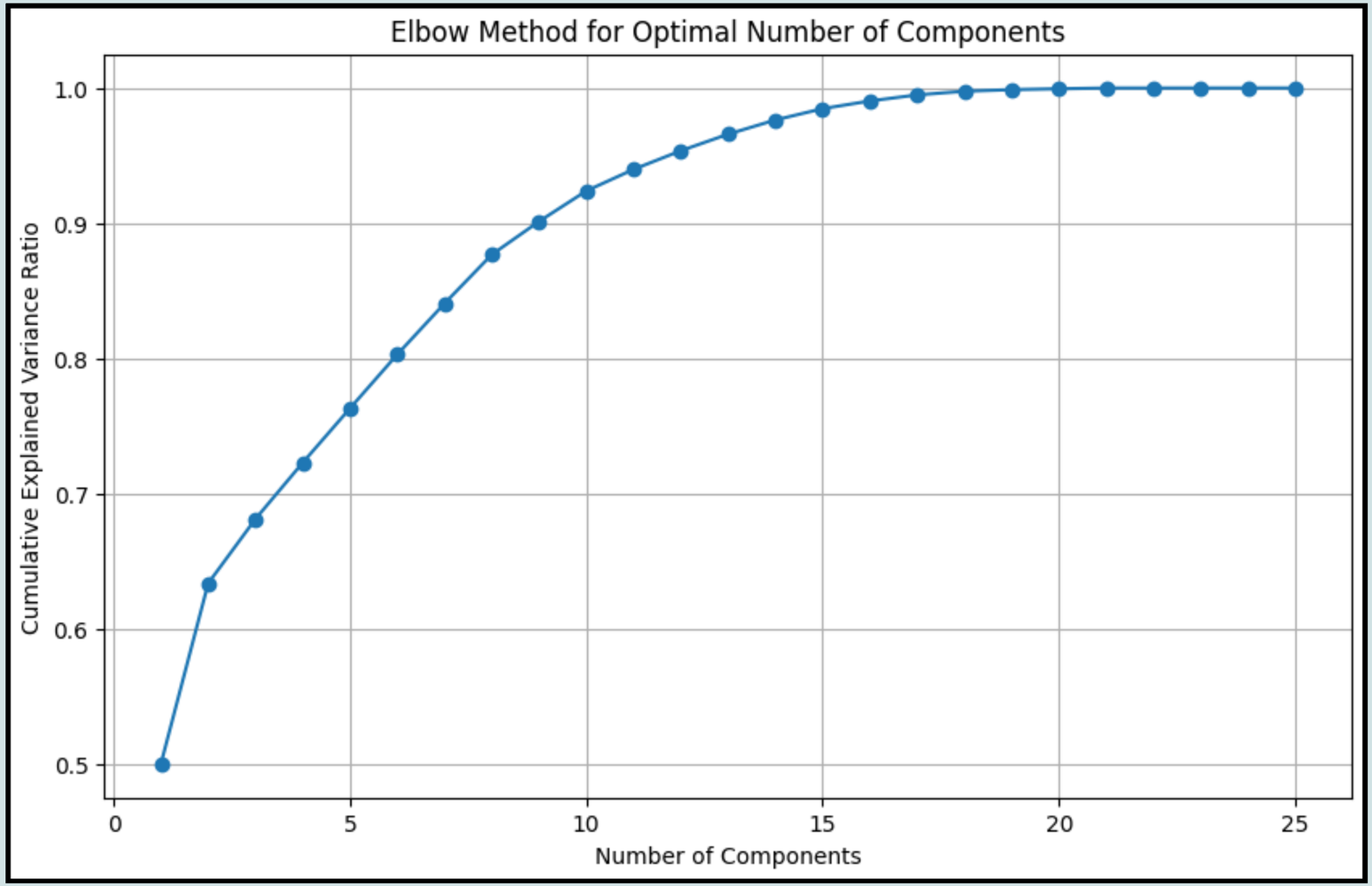
Features:

- Games Played
- Minutes Played
- Field Goals
- Assists
- Rebounding_Impact

Rebounding Impact replaces the rebounds

PCA

PCA Elbow graph to find the best components

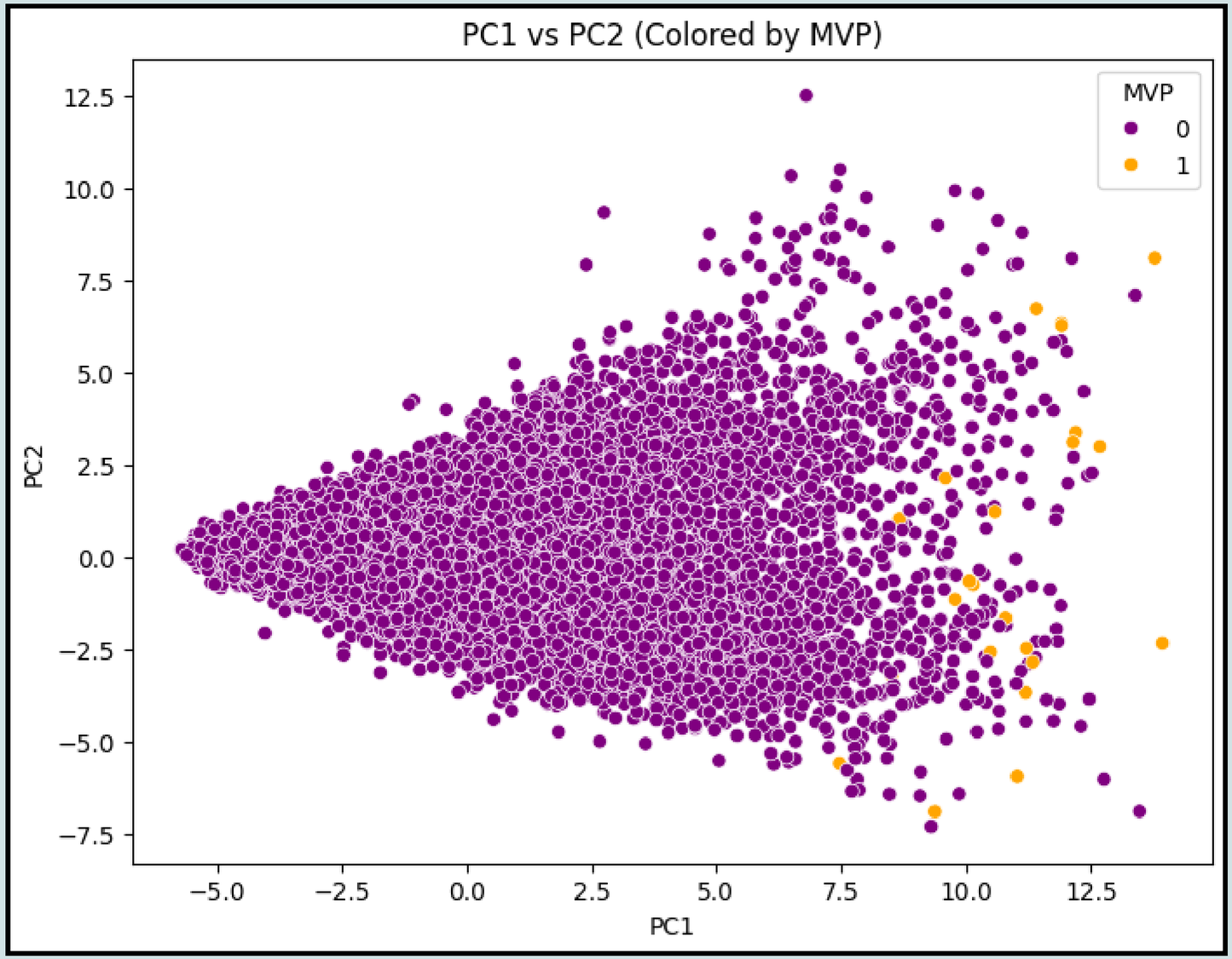


Conclusion:

- An apparent elbow point emerges around the `3-5-7` component range.

PC1 vs PC2

PC1 vs PC2 colored by MVP

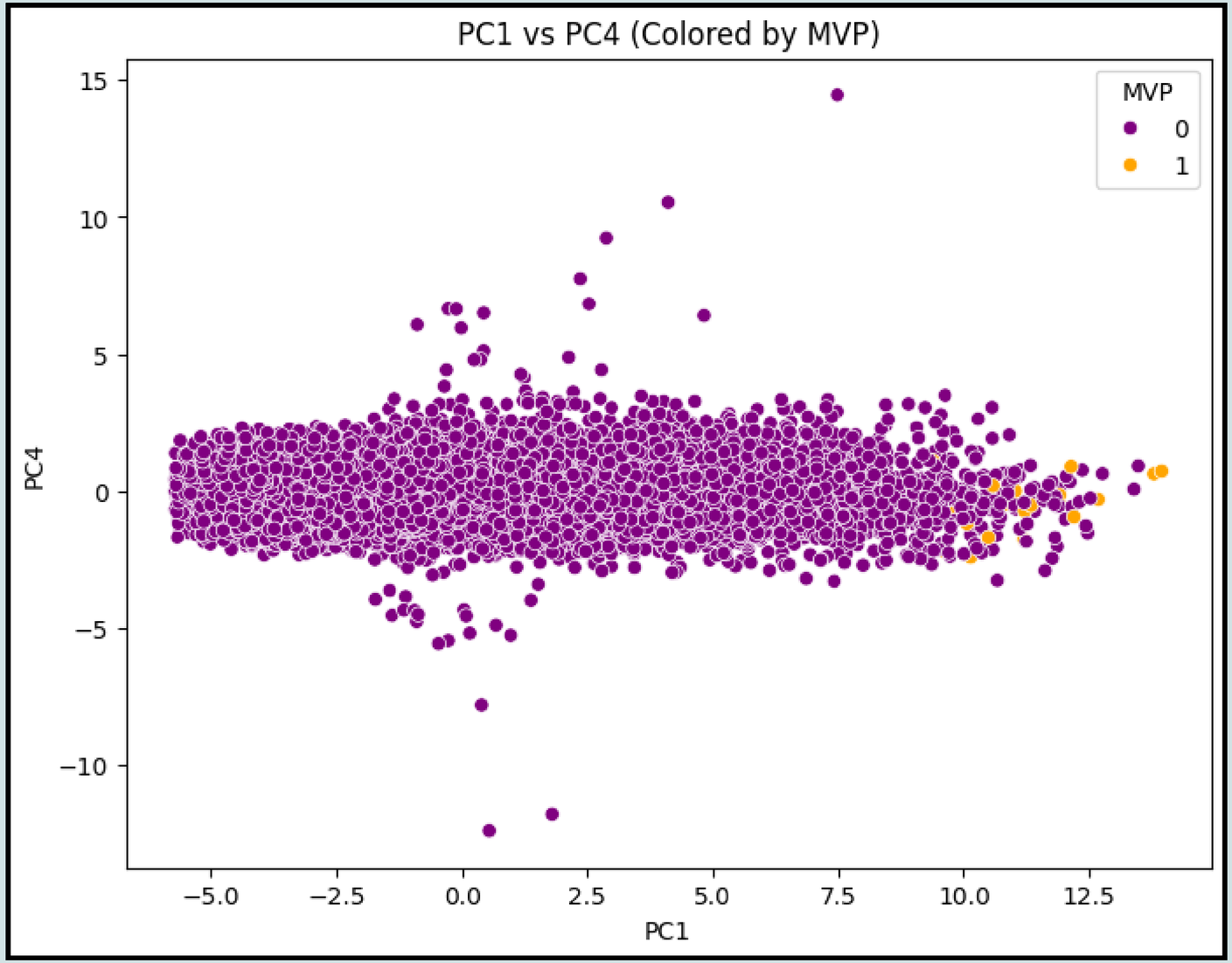


Conclusion:

- **MVPs** predominantly occupy the higher end of the PC1 axis
- **Notable overlap** between MVP and non-MVP players

PC1 vs PC4

PC1 vs PC2 colored by MVP



Conclusion:

- Still **significant overlap** between MVP and non-MVP players

Model Building

Building five models to predict MVP

Class Imbalance:

Over- and undersampling on training and testing.

```
# Applying SMOTEENN for combined over- and under-sampling on the training set
smote_enn = SMOTEENN(random_state=42)
X_train_resampled, y_train_resampled = smote_enn.fit_resample(X_train, y_train)

# Applying SMOTE for oversampling only on the testing set
smote = SMOTE(random_state=42)
X_test_resampled, y_test_resampled = smote.fit_resample(X_test, y_test)
```

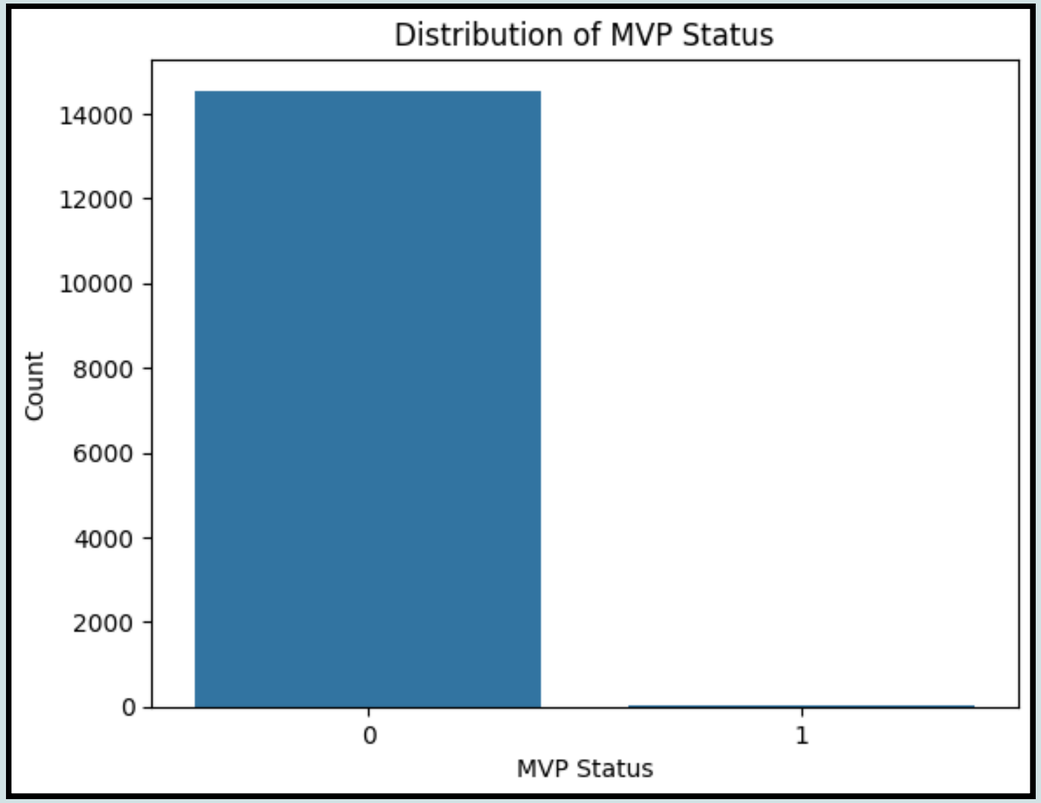
5 Models Created:

- Logistic Regression
- SVM Classifier
- KNN Classifier
- Random Forest Classifier
- Naive Bayes Classifier

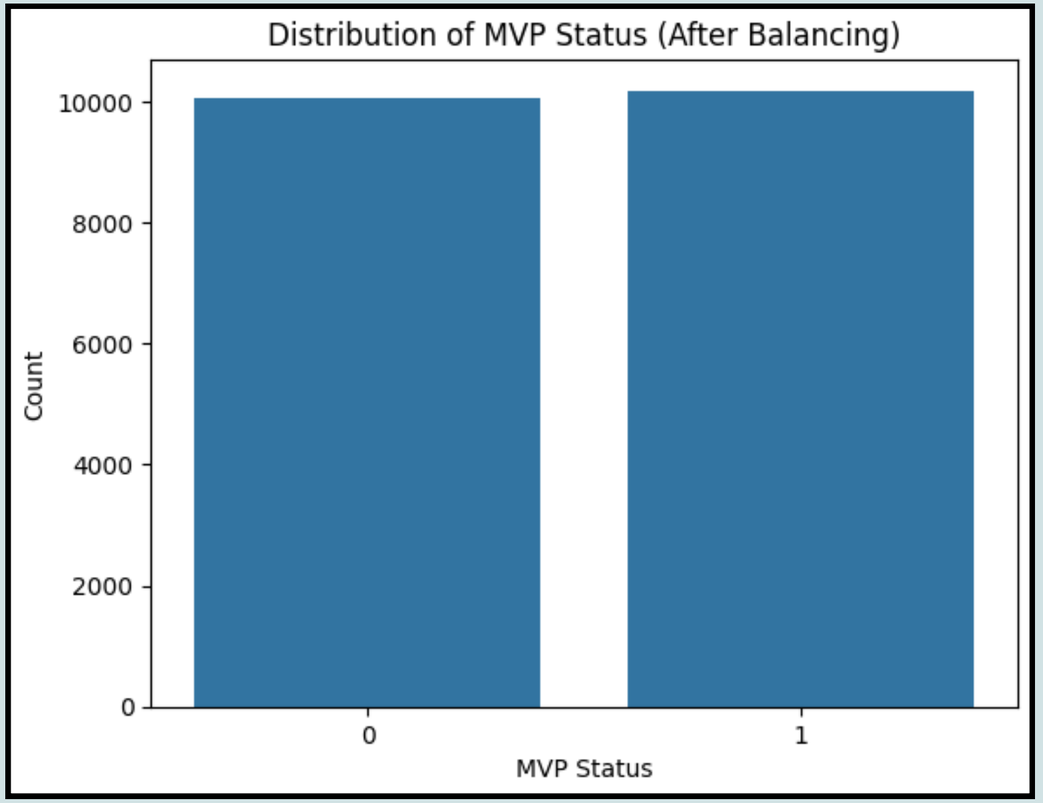
```
# Initialize Logistic Regression classifier
logistic_classifier = LogisticRegression(random_state=42)

# Initialize SVM classifier
svm_classifier = SVC(random_state=42)

# Initialize KNN classifier
knn_classifier = KNeighborsClassifier()

# Initialize RandomForest classifier
random_forest_classifier = RandomForestClassifier(random_state=42)

# Initialize GradientBoosting classifier
gradient_boosting_classifier = GradientBoostingClassifier(random_state=42)
```

FIXING CLASS IMBALANCE MVP

BEFORE



AFTER



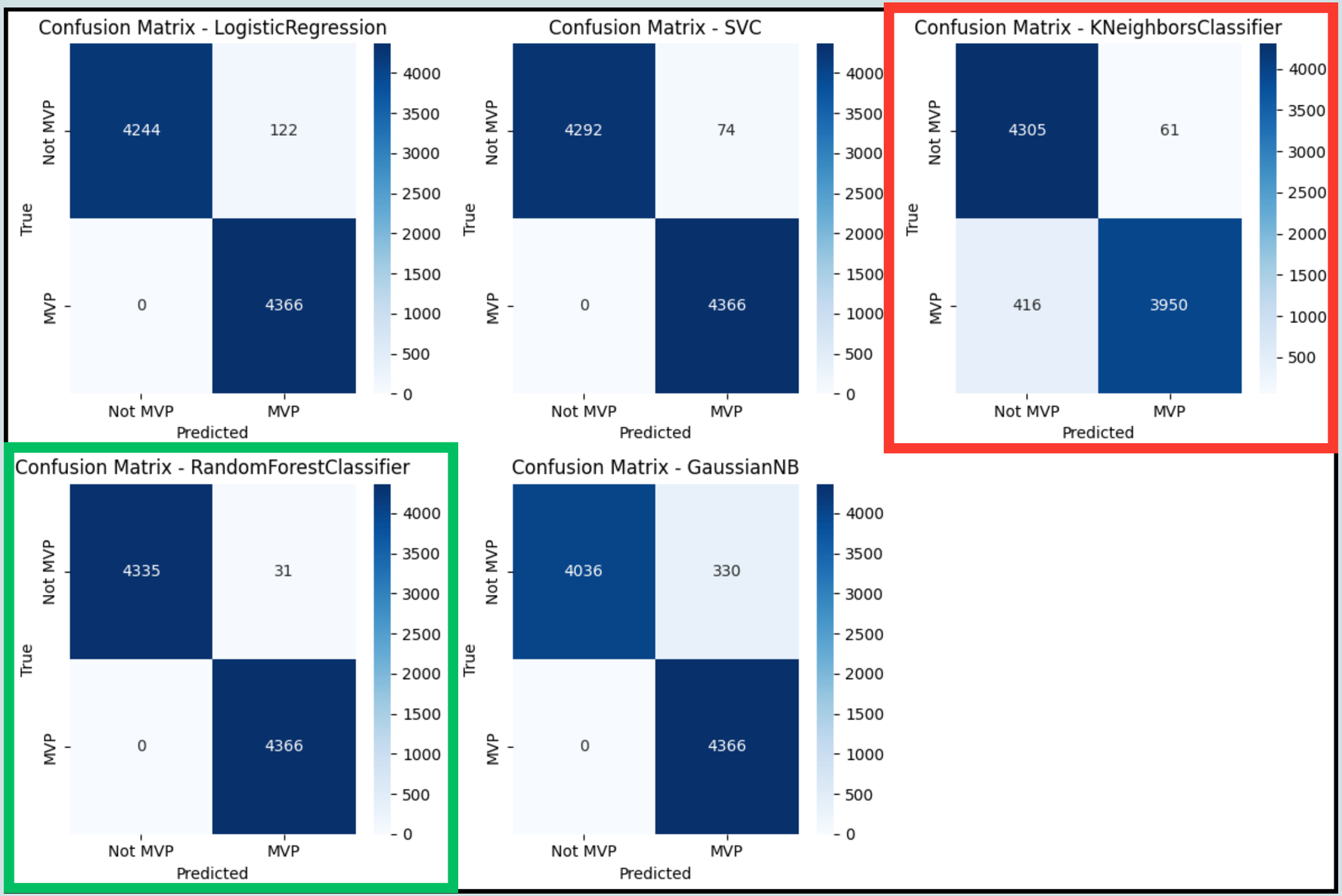
Confusion Matrices

Plotting outcomes of models.

- **Good** Overall Results:
  - Accuracy
  - Precision
  - F1 Score

- **KNN Performs Bad** in terms of Business Case

- **Random Forest** is **Best** Model

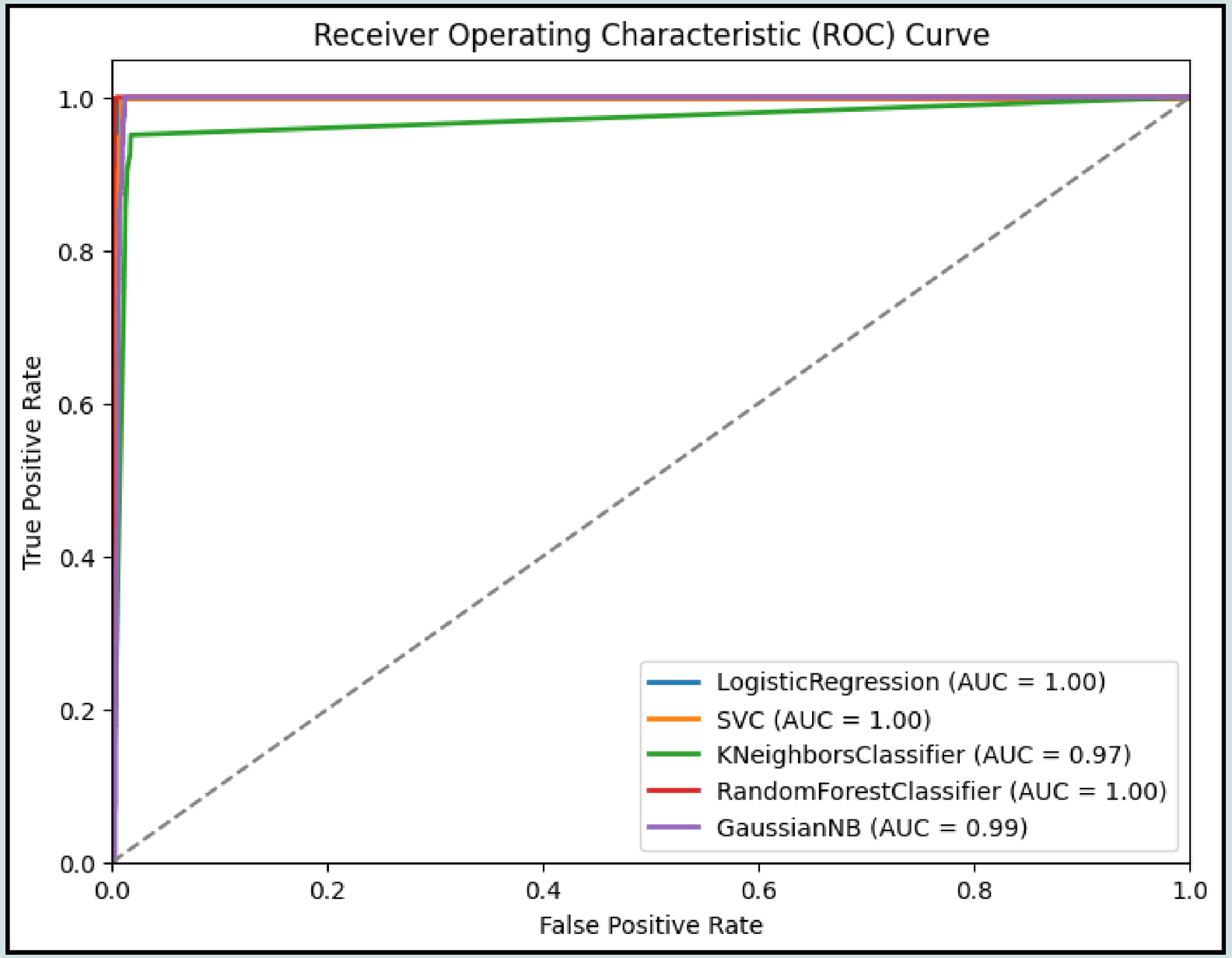


ROC Curve

False positives over true positives.

- **Perfect AUC for:**
  - Logistic Regression
  - SVM Classifier
  - RandomForest Classifier

- **KNN Performs Bad** in terms of Business Case

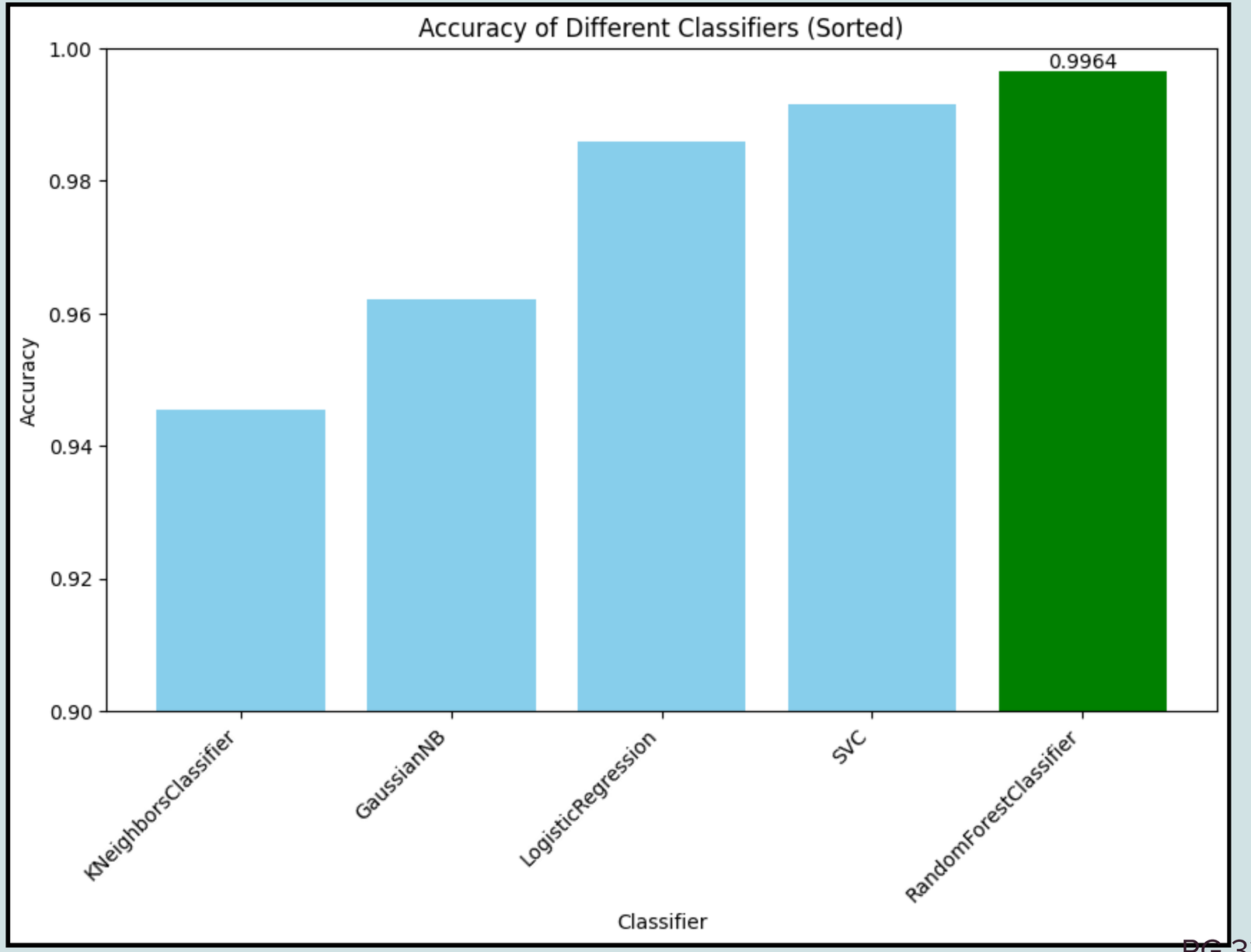


Ranked Accuracy

Worst to best accuracy per model,

- **KNN Performs Worst**
- **Random Forest Performs Best**

= Random Forest as chosen model

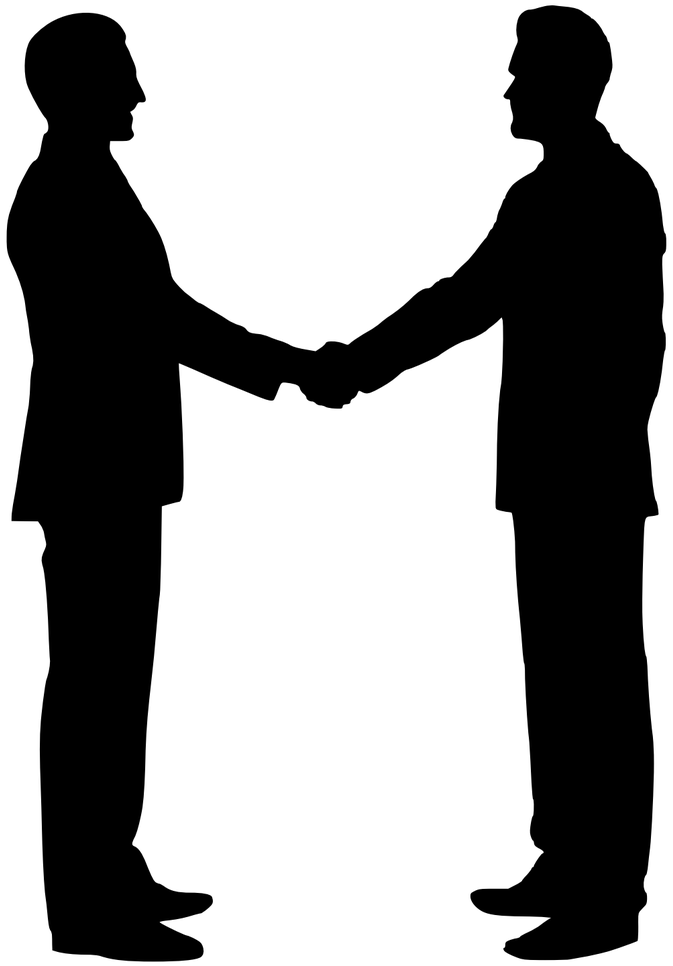


Conclusion

- **Exceeding Business Case Expectations:**
  - Minimal false negatives, showcasing the effectiveness of Random Forest.

- **Optimization of Scouting Processes:**
  - Invaluable Tool for Optimizing Scouting Efforts.
  - Competitive Advantage for Team.
  - Budgetal freedom for the client.

- **Algorithm computationally efficient:**
  - Only 5 out of 23 features.
  - Backed by KBest, LASSO, Decision Trees, and Interaction Terms.
  - PCA Showed no Classification Capabilities.



FIN.

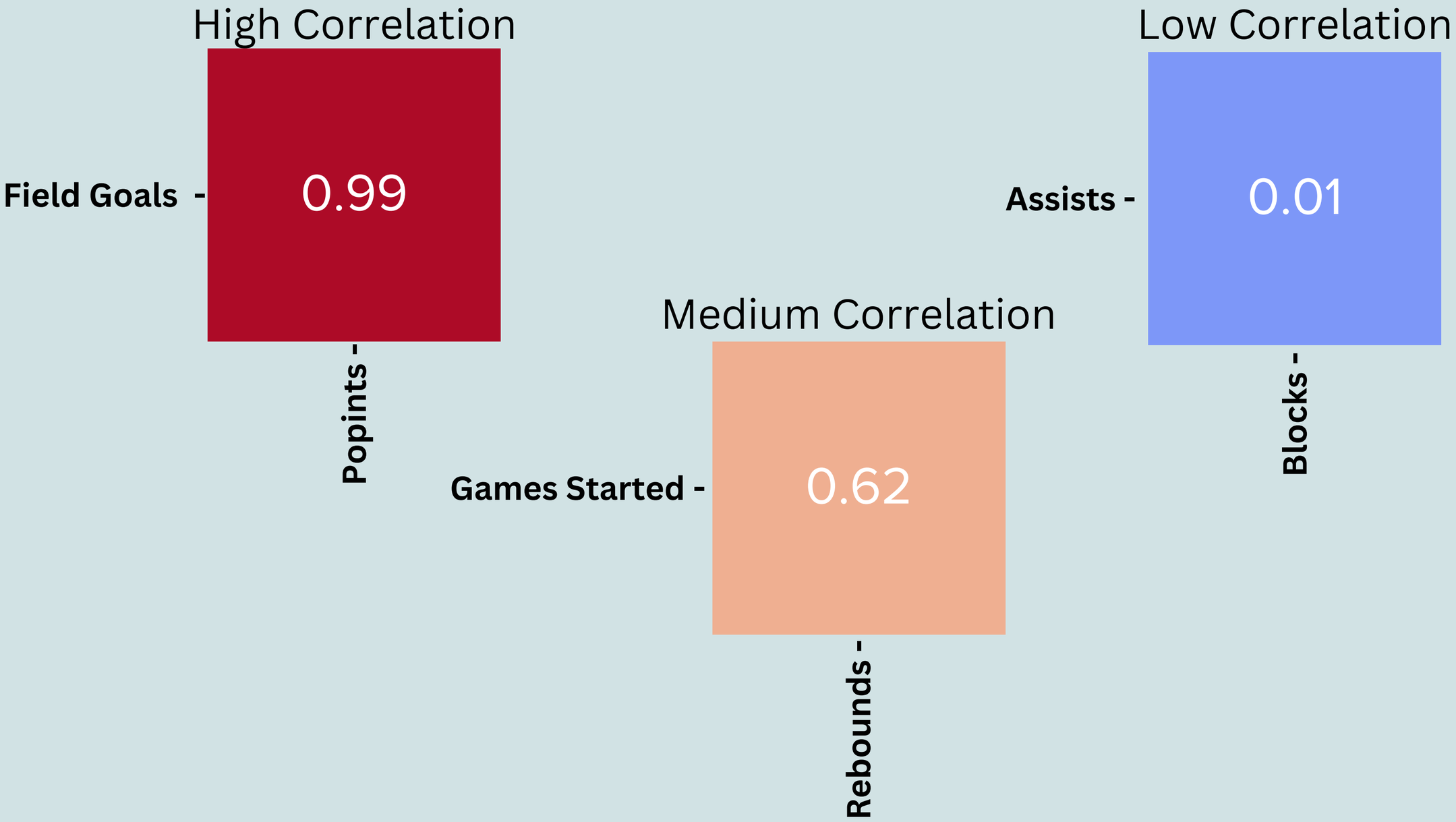
BAN A - Team 3

Vincent, Candela, Esteban, Mehul

OUR PROPOSED SOLUTION

- Establish competitive advantage through enhanced player prediction accuracy, bolstering market positioning.
- Deliver budgetary freedom by minimizing false negatives in MVP-caliber player predictions, and **optimizing resource allocation.**
- Elevate MVP-caliber player predictions to ensure a superior edge in decision-making and team performance.

Correlation Matrix



Feature Selection with Interaction Terms

KBest

K = 5

Features:

- Field Goals
- 2 Pointers
- Free Throws
- Free Throw Attempts
- Points

No Changes

LASSO

Alpha = 0.6

Features:

- Scoring Efficiency
- Free Throws
- Age
- Games Played
- Games Started

Scoring Efficiency Replaces Points Scored.

Decision Tree

Features:

- Games Played
- Minutes Played
- Field Goals
- Assists
- Rebounding_Impact

Rebounding Impact replaces the rebounds